

Benjamin Huang

1-510-589-6864 | benhua476@berkeley.edu | benhuang.org | github.com/benhuang3

EDUCATION

University of California, Berkeley

Berkeley, CA

B.S. in Electrical and Computer Engineering, B.A. in Physics; GPA: 3.91

Aug. 2024 – Present

Relevant Coursework: Microfabrication Architecture, Quantum Computing, Computer Architecture, Data Structures & Algorithms, Linear Algebra, Probability & Statistics, Discrete Mathematics

EXPERIENCE

Machine Learning Engineer

May 2026 – Present

Traffic Patterns / Open Project

Berkeley, CA

- Architecting CloudTCP, a web platform that converts a user placed work zone on aerial imagery into an engineer-quality Traffic Control Plan (TCP) using computer vision and a rules-driven layout engine
- Fine-tuning a Gemini vision-language model on 500+ historical TCPs to generate vectorized traffic control layouts from road geometry and work zone parameters
- Designing a deterministic rules engine that encodes MUTCD and Caltrans spacing, taper, and buffer tables to validate every generated TCP against published standards

FPGA Engineering Intern

May 2026 – August 2026

ChipChop at Skydeck

Berkeley, CA

- Designed an agentic test generation system for chip designs that turns natural language specifications into executable test sequences
- Built custom simulation tools and functional and port coverage skills using IVerilog and reflection loops to iterate on coverage results until the spec is met.
- Increased functional coverage to 97.7% and port coverage to 99.9% for large FPGA designs with 100+ modules

Agentic AI Software Engineer

January 2026 – May 2026

IBM / Open Project

Berkeley, CA

- Designed and deployed a multi-agent system for IBM's internal education platform using IBM Granite, coordinating multiple specialized agents to deliver personalized learning content to IBM engineers
- Architected the agent control flow with tool use, structured output, reflection loops, and retrieval augmented generation (RAG) pipeline for IBM technical documentation

Lead Backend Engineer

January 2026 – May 2026

Behn Meyer / Open Project

Berkeley, CA

- Architected and deployed a FastAPI backend serving REST endpoints for a production agricultural e-commerce mobile app, backed by PostgreSQL with Redis for caching and deployed on AWS EC2 behind an Application Load Balancer
- Implemented OAuth 2.0 authentication and role-based access control via Auth0, securing user sessions and API access across multiple user roles with JWT-based token validation
- Designed horizontal scaling strategy combining Redis caching, database read replicas, and sharding to support 10,000+ concurrent users with sub-200ms p95 API latency

GPU Kernel Optimization Researcher

August 2025 – December 2025

CMU Catalyst Group, Carnegie-Mellon University

Pittsburgh, PA

- Co-developed MPK, a high-performance computing system that fuses GPU operations into a single execution pass, reducing LLM inference latency by 6.7×
- Engineered scalable optimizations for production-scale ML systems, reducing memory overhead and improving computational throughput across distributed workloads
- Implemented fused GPU operators, merging multiple sequential operations into single kernels keeping intermediates on-chip

PROJECTS

dropper.ai | *BrowserUse, Gemini*

March 2026

- Developed dropper.ai, a platform that helps small businesses compete in the e-commerce market through price scraping and website navigation
- Finetuned Gemini agents with BrowserUse to efficiently process and navigate through DOM trees to scrape webpages at scale
- Won Best Solo Hack at DiamondHacks2026

DECODE | *Python, PyTorch, EVO2*

August 2025

- Developed DECODE, a tool that predicts strength of protein-ligand interactions and drug-protein binding efficacy utilizing state-of-the-art transformers and feed-forward neural networks
- Trained a model on the BindingDB Dataset, which includes over 600,000 measured inhibitory effects of drugs on proteins
- Won 2nd place at QBI Hackathon 2025

TECHNICAL SKILLS

Languages: Python, C, C++, CUDA, Java, SQL, Bash, TypeScript, SystemVerilog

Backend & Infrastructure: FastAPI, REST APIs, PostgreSQL, Redis, Docker, AWS (EC2, Application Load Balancer), Nginx, Linux, Git, CI/CD, horizontal scaling (caching, read replicas, sharding), OAuth 2.0, JWT, Auth0

Machine Learning & AI: PyTorch, NumPy, Pandas, transformers, fine-tuning, computer vision, multi-agent systems, tool use, structured output, reflection loops, retrieval-augmented generation (RAG), BrowserUse